

Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP05/005440

International filing date: 17 March 2005 (17.03.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2004-079077
Filing date: 18 March 2004 (18.03.2004)

Date of receipt at the International Bureau: 07 April 2005 (07.04.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)



World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

日 本 国 特 許 庁
JAPAN PATENT OFFICE

17.3.2005

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 4 年 3 月 1 8 日
Date of Application:

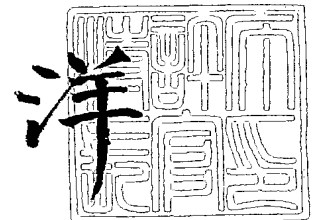
出 願 番 号 特 願 2 0 0 4 - 0 7 9 0 7 7
Application Number:
[ST. 10/C] : [J P 2 0 0 4 - 0 7 9 0 7 7]

出 願 人 日 本 電 気 株 式 有 限 公 司
Applicant(s):

2 0 0 4 年 8 月 3 0 日

特許庁長官
Commissioner,
Japan Patent Office

小 川



【書類名】 特許願
【整理番号】 34403346
【提出日】 平成16年 3月18日
【あて先】 特許庁長官 殿
【国際特許分類】 G06F 17/30
G06F 9/44
G06F 17/27

【発明者】
【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内
【氏名】 坂尾 要祐

【発明者】
【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内
【氏名】 佐藤 研治

【発明者】
【住所又は居所】 東京都港区芝五丁目 7 番 1 号 日本電気株式会社内
【氏名】 赤峯 享

【特許出願人】
【識別番号】 000004237
【氏名又は名称】 日本電気株式会社

【代理人】
【識別番号】 100080816
【弁理士】
【氏名又は名称】 加藤 朝道
【電話番号】 045-476-1131

【手数料の表示】
【予納台帳番号】 030362
【納付金額】 21,000円

【提出物件の目録】
【物件名】 特許請求の範囲 1
【物件名】 明細書 1
【物件名】 図面 1
【物件名】 要約書 1
【包括委任状番号】 9304371

【書類名】 特許請求の範囲**【請求項 1】**

入力した文書から文構造を作成する手段と、
前記文構造の部分構造に対して、予め定められた所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する手段と、
前記意味の類似したパターンを同一パターンと判定してパターン検出を行う手段と、
を備えている、ことを特徴とするテキストマイニング装置。

【請求項 2】

テキストマイニングの対象となる文書の集まりを記憶する記憶部と、
前記記憶部の前記文書を入力して解析し文構造を取得する解析部と、
を備え、
前記解析部は、前記文書を解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する、ことを特徴とする請求項 1 に記載のテキストマイニング装置。

【請求項 3】

前記類似構造を生成する手段が、
前記文構造について並列変形を行う手段と、
前記文構造の部分構成を生成する手段と、
前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、
同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、
前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、
を備え、
前記類似構造を前記文構造の部分構造の同値類とする、ことを特徴とする、請求項 1 に記載のテキストマイニング装置。

【請求項 4】

テキストマイニングの対象となる文書の集まりを記憶する記憶部と、
前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、
前記解析部により解析して得られる文構造の部分構造に対して予め定められた所定の変形操作を行い、意味的に類似したパターン of 類似構造を生成する類似構造生成部と、
前記類似構造生成部によって生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行うパターン検出部と、
を備えている、ことを特徴とするテキストマイニング装置。

【請求項 5】

前記パターン検出部は、前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出する、ことを特徴とする請求項 4 に記載のテキストマイニング装置。

【請求項 6】

前記類似構造生成部が、
前記文構造について並列変形を行う手段と、
前記文構造の部分構成を生成する手段と、
前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、
同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、
前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、
を備え、
前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項 4 に記載のテキストマイニング装置。

【請求項 7】

使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する手段を備えている、ことを特徴とする請求項 4 に記載のテキストマイニング装置。

【請求項 8】

テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、
使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第 1 の指定項目を生成する類似構造生成調整部と、
使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第 2 の指定項目を生成する類似構造判定調整部と、
前記類似構造生成調整部によって生成された第 1 の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する類似構造生成部と、
前記類似構造生成部によって生成された類似構造を生成元の部分構造の同値類として扱い、前記類似構造判定調整部の第 2 の指定項目に従い、属性値の差異を無視しながら、頻出パターンの検出を行う類似パターン検出部と、
を備えている、ことを特徴とするテキストマイニング装置。

【請求項 9】

前記解析部は、前記文書を解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした前記文構造を生成し、

前記属性値は、前記文構造に付加された表層格及び／又は付属語情報を含む、ことを特徴とする請求項 8 に記載のテキストマイニング装置。

【請求項 10】

前記類似パターン検出部は、頻出の類似パターンを検出する、ことを特徴とする請求項 8 に記載のテキストマイニング装置。

【請求項 11】

前記類似構造生成部が、

前記第 1 の指定項目に、並列変形の指定がある場合、前記文構造について並列変形を行う手段と、

前記文構造の部分構造を生成する手段と、

前記第 1 の指定項目に、有向枝の無向枝化の指定がある場合に、前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、

前記第 1 の指定項目に、同義語の置換の指定がある場合、同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、

前記第 1 の指定項目に、順序木の無順序木化の指定がある場合、前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、

を備え、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項 8 に記載のテキストマイニング装置。

【請求項 12】

入力した文書から文構造を作成する工程と、

前記文構造の部分構造に対する所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する工程と、

前記意味の類似したパターンを同一パターンと判定してパターン検出を行う工程と、

を含む、ことを特徴とするテキストマイニング方法。

【請求項 13】

テキストマイニングの対象となる文書の集まりを記憶する記憶部から前記文書を入力して解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する工程を含む、ことを特徴とする請求項 12 に記載のテキストマイニング方法。

【請求項 14】

前記類似構造を生成する工程が、

前記文構造について並列変形を行う工程と、

前記文構造の部分構造を生成する工程と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、
前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、
を含み、
前記類似構造を前記部分構造の同値類とする、ことを特徴とする請求項 12 に記載のテキストマイニング方法。

【請求項 15】

テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、

前記文構造の部分構造に対して予め定められた所定の変形操作を行い、意味的に類似したパターンを有する類似構造を生成する工程と、

前記生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行う工程と、

を含む、ことを特徴とするテキストマイニング方法。

【請求項 16】

前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出する工程を含む、ことを特徴とする請求項 15 に記載のテキストマイニング方法。

【請求項 17】

前記類似構造を生成する工程が、

前記文構造について並列変形を行う工程と、

前記文構造の部分構造を生成する工程と、

前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、

前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、

を含み、

前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項 15 に記載のテキストマイニング方法。

【請求項 18】

使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する工程を備えている、ことを特徴とする請求項 17 に記載のテキストマイニング方法。

【請求項 19】

テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、

使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第 1 の指定項目を生成する工程と、

使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第 2 の指定項目を生成する工程と、

前記生成された第 1 の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変形操作を行い、前記部分構造と意味的に類似した類似構造を生成する工程と、

前記生成された類似構造を生成元の部分構造の同値類として扱い、前記第 2 の指定項目に従い、属性値の差異を無視してパターンの検出を行う工程と、

を含む、ことを特徴とするテキストマイニング方法。

【請求項 20】

前記文構造を取得する工程が、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした前記文構造を生成し、

前記属性値は、前記文構造に付加された表層格及び／又は付属語情報を含む、ことを特徴とする請求項 19 に記載のテキストマイニング方法。

【請求項 21】

前記頻出の類似パターンを検出する、ことを特徴とする請求項 19 に記載のテキストマイニング方法。

【請求項 22】

前記類似構造を生成する工程が、
前記第1の指定項目に、並列変形の指定がある場合、前記文構造について並列変形を行う工程と、
前記文構造の部分構造を生成する工程と、
前記第1の指定項目に、有向枝の無向枝化の指定がある場合に、前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、
前記第1の指定項目に、同義語の置換の指定がある場合、同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、
前記第1の指定項目に、順序木の無順序木化の指定がある場合、前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、
を含み、
前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項19に記載のテキストマイニング方法。

【請求項23】

テキストマイニング装置を構成するコンピュータに、
テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、
前記文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する処理と、
前記生成された類似構造を、生成元の部分構造の同値類として扱い、所定のパターン検出を行う処理と、
を実行させるプログラム。

【請求項24】

テキストマイニング装置を構成するコンピュータに、
テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、
前記文構造の部分構造に対して予め定められた所定の変換操作を行い、前記部分構造と意味的に類似したパターンの類似構造を生成する処理と、
前記生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異を無視しながらパターン検出を行う処理と、
を実行させるプログラム。

【請求項25】

テキストマイニング装置を構成するコンピュータに、
テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、
使用者の入力から、前記文構造の差異の種別ごとに、同一構造と判定するか否かを指定する第1の指定項目と、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第2の指定項目を生成する処理と、
前記文構造の差異の種別ごとに同一構造と判定するか否かを指定する前記第1の指定項目に従って、前記文構造の部分構造に対して所定の変換操作を行い、意味的に類似したパターンの類似構造を生成する処理と、
生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異の種別ごとに同一構造と判定するか否かを指定する前記第2の指定項目に従って、属性値の差異を無視しながら頻出パターンの検出を行う処理と
を実行させるプログラム。

【書類名】明細書

【発明の名称】テキストマイニング装置、その方法及びプログラム

【技術分野】

【0001】

本発明は、構文解析などを用いて、コンピュータ上に蓄積される電子化テキストを構造化して分析を行うテキストマイニング装置、テキストマイニング方法及びテキストマイニング用プログラムに関し、特に、意味の類似した文の構造を同一の構造と判定して分析を行うことができるテキストマイニング装置、テキストマイニング方法及びテキストマイニング用プログラムに関する。

【背景技術】

【0002】

テキストマイニング装置の一例として、図10に示すような構成が知られている（後記特許文献1参照）。図10に示すように、この従来のテキストマイニング装置は、基本辞書記憶部と、文書データ記憶部と、分野依存辞書記憶部と、言語特徴分析装置と、言語解析装置と、パターン抽出装置と、頻出パターン表示装置とを備えている。

【0003】

図10に示した従来のテキストマイニング装置は、概略、つぎのように動作する。まず、言語特徴分析装置によって基本辞書と文書データとから分野依存辞書を作成し、言語解析装置によって基本辞書と分野依存辞書と文書データから構文木等の構造を作成し、パターン抽出装置が、この構造を用いて頻出パターンを抽出し、この頻出パターンに合致する文書データ中の文書を、頻出パターン適合文書記憶部に記憶させると同時に、この頻出パターンを出力する。

【0004】

一般的に、言語解析装置によって作成される構造として、例えば、

(A1) 文中の文節を、構造の節点で表し、

(A2) 付属語情報を、節点の属性値で表し、

(A3) 係り受け関係を、係り元の節点から係り先の節点への有向枝で表し、

(A4) 表層格の情報を、有向枝の属性値で表す

という構造が良く用いられる。

【0005】

ここで、付属語情報とは、付属語によって文節に付加される、進行や完了などの時制、容易や困難などのモダリティ、及び否定などの付属的な概念である。図11に、この形式で表された「彼は車種Aが価格を下げたのを知らない」という文の構文構造の一例を示す。文の文節、「彼」、「車種A」、「価格」、「下げる」、「知る」を節点で表わし、付属語情報を、節点の属性値で表し（節点「知る」の属性値として、付属語情報：否定）、係り受け関係を、係り元の節点から係り先への有向枝で表わし（例えば「彼」→「知る」）、表層格の情報を、有向枝の属性値で表している（例えば「彼」→「知る」の有向枝の属性値として「表層格：は」）。

【0006】

また、構造中のこれらの情報は、全て属性値を持たないラベル付きの節点と、属性値を持たない有向枝のみからなる構造で表現することも可能である。図12に、この形式で表された「彼は車種Aが価格を下げたのを知らない」という文の構文構造の例を示す。文の文節「彼」、「車種A」、「価格」、「下げる」、「知る」を属性値を持たないラベル付きの節点で表わし（例えば節点「彼」には「表層格：は」のラベルが付加され、「下げる」には、ラベル「付属語情報：完了」、「表層格：を」が付加されている）、係り元の節点から係り先への有向枝は属性値を持たない有向枝とされる。

【0007】

【特許文献1】特開2001-84250号公報（第4、5頁、第3図）

【発明の開示】

【発明が解決しようとする課題】

【0008】

上記した従来のシステムは下記記載の問題点を有している。なお、以下の問題点及びその解析は、本願発明者らによる研究・検討結果に基づくものであり、図13、図14の内容は、問題の在り処を具体的に説明するために、本願発明者らが提示したものである。

【0009】

第1の問題点は、頻出パターン検出の際に、意味は類似しているが、連結構造が異なる構造を、全く別のパターンと判定してしまう、ということである。

【0010】

連結構造とは、構造の節点と単語文字列及び有向枝の連結関係と、向きにのみ注目し、付属的な属性情報を省略した構造のことをいう。

【0011】

上記第1の問題点が生じる理由は、従来のテキストマイニング装置は、連結構造が異なり、類似した意味を持つ構造を同一と判定する手段を具備していないためである。

【0012】

属性値を用いた文構造を用いる際に、連結構造が異なり、類似した意味を持つ構造の差異の例として、

- (B1) 係り受けの向きの差異、
- (B2) 係り受けの順序の差異、
- (B3) 同義語の置換による差異、及び、
- (B4) 並列の構文構造と意味構造の差異

などが挙げられる。

【0013】

図13に、これらの連結構造による構造の差異の例を示す。また、属性値を用いない文構造を用いる際は、あらゆる意味の類似した構造の差異は、連結構造の差異で表現される。

【0014】

図13(a)に示す例では、意味の類似した「速いのは車種A」と「車種Aは速い」の連結構造において、係り元と係り先が相違している。

【0015】

図13(b)に示す例では、意味の類似した「速く安い車種A」と「安く速い車種A」の連結構造において、係り元の「速い」と「安い」の節点の順序関係が、相違している。

【0016】

図13(c)に示す例では、意味の類似した「車種Aは速い」と「車種Aは高速だ」のそれぞれの連結構造において、係り先の「速い」と「高速」が相違している。

【0017】

図13(d)に示す例では、「車種Aと車種Bは速い」の構文構造と意味構造を表わしている。図3(d)において、係り元「車種A」が「車種B」に係り「車種B」が「速い」に係る連結構造と、係り元「車種A」と「車種B」から係り先「速い」への有向枝を有する連結構造がある。

【0018】

第2の問題点は、頻出パターン検出の際に異なる属性値を持つが、類似した意味を持つ構造を、全く別のパターンと判定してしまう、ということである。

【0019】

その理由は、従来のテキストマイニング装置では、異なる属性値を持つ構造を、同一と判定することについて、何ら考慮されていないためである。

【0020】

属性値を用いた文構造を用いる際に、属性値が異なり、類似した意味を持つ構造の差異の例として、付属語情報の差異、表層格の差異などが挙げられる。図14に、これらの属性値による構造の差異の例を示す。

【0021】

図14(a)に示す例では、類似した意味を持つ「車種Aは加速」と「車種Aの加速」の連結構造において、有向枝の表層格が相違している。

【0022】

図14(b)に示す例では、類似した意味を持つ「車種Aは速い」と「車種Aは速かった」の連結構造において、係り先の節点「速い」の付属語情報が相違している。

【0023】

第3の問題点は、テキストマイニング装置の使用者（ユーザ）がどこまで類似した構造を同一な構造と判定して頻出パターンの検出を行うかを調整できない、ということである。

【0024】

その理由は、従来のテキストマイニング装置では、使用者が頻出パターン検出の際にどのような構造を同一と判定するかを調整することについて、何ら考慮されていないためである。

【0025】

したがって、本発明の目的は、類似した意味を持ち、連結構造の異なる構造を、同一のパターンと判定して頻出パターン等の検出を行うテキストマイニング装置及び方法並びにプログラムを提供することにある。

【0026】

本発明の他の目的は、類似した意味を持ち属性値の異なる構造を同一な構造と判定して頻出パターン検出を行うかを調整できるテキストマイニング装置及び方法並びにプログラムを提供することにある。

【0027】

本発明のさらに他の目的は、テキストマイニングの使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整できるテキストマイニング装置及び方法並びにプログラムを提供することにある。

【課題を解決するための手段】

【0028】

本願で開示される発明は、上記目的を達成するため、概略以下の構成とされる。

【0029】

本発明の第1のAspect（側面）に係るテキストマイニング装置は、入力した文書から文構造を作成する手段と、前記文構造の部分構造に対して予め定められた所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する手段と、前記意味の類似したパターンを同一パターンと判定してパターン検出を行う手段と、を備えている。

【0030】

本発明において、前記類似構造を生成する手段は、前記文構造について並列変形を行う手段と、前記文構造の部分構造を生成する手段と、前記文書構造及び／又は部分構造の有向枝の無向枝化を行う手段と、同義語辞書を参照して前記文書構造及び／又は部分構造中の同義語の置換を行う手段と、前記文書構造及び／又は部分構造における順序木の無順序木化を行う手段と、を備え、前記類似構造を前記部分構造の同値類とする。同値類とは、構造の集合でその各要素を同一の構造として扱うものをいい、二つの同値類に一つでも、同一の要素が含まれる時には、その二つの同値類を同一の同値類と判定する。本発明によれば、生成された類似構造を生成元の文構造の同値類として扱い、頻出パターン検出を行う。

【0031】

本発明の第2のAspect（側面）に係るテキストマイニング装置は、第1のAspectに係るテキストマイニング装置の構成に含まれる頻出パターン検出手段に代わり、構造中の属性値の差異を無視して、頻出パターンの検出を行う頻出類似パターン検出手段を備え、属性値の異なる類似した構造を同一な構造と判定して頻出パターンの検出を行う。本発明によれば、構造中の属性値が異なる類似した構造を同一と判定して頻出パターン検出

を行う。

【0032】

本発明の第3のAspect（側面）に係るテキストマイニング装置は、テキストマイニングの対象となる文書の集まりを記憶する記憶部と、前記記憶部の前記文書を解析して文構造を取得する解析部と、使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第1の指定項目を生成する類似構造生成調整部と、使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第2の指定項目を生成し類似構造判定調整部と、前記類似構造生成調整部によって生成された第1の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する類似構造生成部と、前記類似構造生成部によって生成された類似構造を生成元の部分構造の同値類として扱い、前記類似構造判定調整部の第2の指定項目に従い、属性値の差異を無視しながら、頻出パターンの検出を行う類似パターン検出部と、を備えている。本発明によれば、構造の同一性の判定を調整するための指定の入力を受け付ける。

【0033】

本発明のさらに他のAspect（側面）に係る方法は、
入力した文書から文構造を作成する工程と、
前記文構造の部分構造に対する所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する工程と、
前記意味の類似したパターンを同一パターンと判定してパターン検出を行う工程と、を含む。

【0034】

本発明のさらに他のAspect（側面）に係る方法は、テキストマイニングの対象となるテキストの集まりを記憶する記憶部のテキストを解析して文構造を取得する工程と、
前記文構造の部分構造に対して意味的に類似したパターンの類似構造を生成する工程と、
生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異を無視しながらパターンの検出を行う工程と、を含む。

【0035】

本発明のさらに他のAspect（側面）に係る方法は、テキストマイニングの対象となるテキストの集まりを記憶する記憶部のテキストを解析して文構造を取得する工程と、
入力装置から入力された使用者の入力情報から、文構造（連結構造）の差異の種別ごとに同一構造と判定するか否かを指定する第1の指定項目と、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第2の指定項目を生成するステップと、
文構造（連結構造）の差異の種別ごとに同一構造と判定するか否かを指定する第1の指定項目に従い、前記文構造の部分構造に対して意味的に類似した構造を生成する工程と、
生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第2の指定項目に従い、属性値の差異を無視しながら頻出パターンの検出を行う工程と、を含む。

【0036】

本発明のさらに他のAspect（側面）に係るプログラムは、テキストマイニング装置を構成するコンピュータに、
テキストマイニングの対象となるテキストの集まりを記憶する記憶部の前記テキストを解析して文構造を取得する処理と、
前記処理で解析して得られた文構造の部分構造に対して、意味的に類似した構造を生成する処理と、
生成された類似構造を、生成元の部分構造の同値類として扱い、頻出パターンの検出を行う処理と、
を実行させるプログラムよりなる。

【発明の効果】

【0037】

本発明によれば、連結構造は異なるが類似した意味を持つ構造を同一の構造と判定して頻出パターンを検出することができる。本発明によれば、属性値を持たない構造の集合に対して類似構造を同一と判定して頻出パターンの検出を行うことができる。

【0038】

その理由は、本発明においては、生成した類似構造を元の構造の同値類として扱い、頻出パターン検出を行う構成としたためである。本発明によれば、属性値を持つ構造の集合に対しても類似構造を同一と判定して頻出パターンの検出を行うことができる。

【0039】

また、本発明によれば、類似した意味を持つが異なる属性値を持つ構造を同一の構造と判定して頻出パターンを検出することができる。

【0040】

その理由は、本発明においては、頻出類似パターン検出手段が属性値の差異を無視して頻出パターン検出を行うためである。

【0041】

さらに本発明によれば、テキストマイニング装置の使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整することができる。

【0042】

その理由は、本発明においては、類似構造生成調整手段と類似構造判定調整手段が使用者からの入力に基づき、どこまで類似した構造を同一な構造と判定するかの調整を行う構成としたためである。

【発明を実施するための最良の形態】**【0043】**

次に、発明を実施するための最良の形態について図面を参照して詳細に説明する。

【0044】

図1を参照すると、本発明を実施するための最良の一形態（第1の実施の形態）は、情報を記憶する記憶装置1と、プログラム制御により動作するデータ処理装置2と、検出されたパターンを出力する出力装置3と、を備えている。記憶装置1はテキストデータベース(DB)11を含む。テキストDB11は、テキストマイニングの対象となるテキストの集合を記憶している。

【0045】

データ処理装置2は、言語解析手段21と、類似構造生成手段22と、頻出パターン検出手段23を含む。これらの手段はそれぞれ概略つぎのように動作する。

【0046】

言語解析手段21は、テキストDB11からテキスト集合を読み込み、集合中の各テキストを解析して文構造を得る。

【0047】

類似構造生成手段22は、言語解析手段21から送られた文構造の集合中の各文構造の全ての部分構造を抽出し、各部分構造の全ての類似構造を生成して類似構造と生成元の部分構造を同値類とする。

【0048】

頻出パターン検出手段23は、類似構造生成手段22から送られた部分構造の同値類の集合から頻出するパターンを検出し、出力装置3へ送る。

【0049】

図2は、本実施形態の動作を説明するための流れ図である。次に、図1及び図2を参照して、本発明の第1の実施形態の動作について詳細に説明する。

【0050】

まず、言語解析手段21が、テキストDB11から、テキスト集合を読み込む。言語解析手段21は、テキスト集合中の各テキストに対し解析を行い、解析結果として、文構造を生成し、類似構造生成手段22に送る（図2のステップA1）。

【0051】

次に、類似構造生成手段 22 は、与えられた文構造の集合中の部分構造の全ての類似構造を生成し、類似構造を生成元の部分構造の同値類とし、同値類の集合を頻出パターン検出手段 23 に送る（図 2 のステップ A2）。

【0052】

さらに、頻出パターン検出手段 23 は、与えられた部分構造の同値類から、頻出パターンの検出を行う（図 2 のステップ A3）。

【0053】

頻出パターン検出手段 23 は、検出した頻出パターンを出力装置 3 に出力する（図 2 のステップ A4）。

【0054】

図 3 は、図 2 のステップ A2 における、類似構造生成手段 22 の動作の詳細なフローチャートを示す図である。

【0055】

図 3 を参照すると、類似構造生成手段 22 は、まず並列構文の構文構造と意味構造の違いに対応するための「並列の変形」を行う（図 3 のステップ A2-1）。

【0056】

次に、文構造全体だけではなく部分構造からもパターン検出を行うための「部分構造の生成」を行う（図 3 のステップ A2-2）。

【0057】

次に、係り受けの向きの差異に対応するための「有向枝の無向枝化」を行う（図 3 のステップ A2-3）。

【0058】

次に、同義語の差異に対応するための「同義語の置換」を行う（図 3 のステップ A2-4）。

【0059】

その係り受けの順序の違いに対応するための「順序木の無順序木化」を行う（図 3 のステップ A2-5）。

【0060】

最後に、類似構造を生成元の部分構造の同値類の要素とすることで、「同値類の生成」を行う（図 3 のステップ A2-6）。

【0061】

次に、本発明の第 1 の実施の形態の作用効果について説明する。

【0062】

本発明の第 1 の実施の形態では、類似構造生成手段 22 が生成した類似構造を、元の構造の同値類として扱い、頻出パターン検出を行うように構成されている。このため、連結構造は異なるが、類似した意味を持つ構造を、同一の構造と判定して、頻出パターンを検出できる。

【0063】

次に、本発明を実施するための最良の別の形態（第 2 の実施の形態）について図面を参照して詳細に説明する。

【0064】

図 4 を参照すると、本発明の第 2 の実施の形態においては、データ処理装置 4 が、図 1 に示された第 1 の実施の形態におけるデータ処理装置 2 の頻出パターン検出手段 23 に代えて、頻出類似パターン検出手段 24 を備えている。言語解析手段 21、類似構造生成手段 22 は、前記第 1 の実施の形態のものと同一である。

【0065】

本発明の第 2 の実施の形態において、頻出類似パターン検出手段 24 は、類似構造生成手段 22 から送られた部分構造の同値類の集合から、属性値の相違を無視しながら、頻出パターンの検出を行い、検出した頻出パターンを出力装置 3 に送る。

【0066】

図5は、本発明の第2の実施形態の動作を説明するための流れ図である。次に、図4及び図5を参照して、本発明の第2の実施形態の動作について詳細に説明する。本発明の第2の実施形態においては、図2のステップA3のかわりに、ステップB3が実行される。図5のステップA1、A2、A4で示される処理は、前記第1の実施の形態の処理と同一であるため、説明は省略する。

【0067】

前記第1の発明の実施の形態では、頻出パターン検出手段23は、連結構造が同一でも属性値の異なる構造は同一と判定せずに、頻出パターンの検出を行っていた。

【0068】

本実施の形態では、頻出類似パターン検出手段24は、類似構造生成手段22から与えられた同値類の集合を、連結構造が同一で属性値の異なる構造も同一な構造と判定しながら頻出パターンの検出を行い、検出された頻出パターンを出力装置3に送る（図5のステップB3）。

【0069】

次に、本発明の第2の実施形態の作用効果について説明する。

【0070】

本発明の第2の実施の形態では、頻出類似パターン検出手段24が、連結構造は同一で属性値の異なる構造も同一な構造と判定しながら頻出パターンの検出を行うように構成されている。このため、意味は類似しているが属性値の異なる構造も同一な構造と判定して頻出パターンの検出を行うことができる。

【0071】

次に、本発明の第3の発明を実施するための最良の形態について図面を参照して詳細に説明する。

【0072】

図6を参照すると、本発明の第3の実施の形態は、図4に示した前記第2の実施の形態の構成と比較し、入力装置6を備え、データ処理装置5が、類似構造生成調整手段25及び類似構造判定調整手段26を備えている。

【0073】

入力装置6は、使用者から、

- ・ 文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、
- ・ 属性値の種別ごとに値の差異を無視するか否かを指定するための入力と、

を受け付け、それぞれを、類似構造生成調整手段25と類似構造判定調整手段26に送る。

【0074】

入力装置6で受け付ける指定の入力の例としては、

- ・ 「使用者から文構造の差異の種別ごとに同一構造と判定するか否かと属性値の種別ごとに値の差異を無視するか否かについての指定項目」、
- ・ 「頻出パターン検出の際に同一パターンを持っていると判定しない文の例」、
- ・ 「頻出パターン検出の際に同一パターンを持っていると判定する文の例」

などが挙げられる。

【0075】

類似構造生成調整手段25は、入力装置6から与えられた指定から、連結構造の差異の種別ごとに同一構造と判定するか否かを決定し、その指定項目を、類似構造生成手段22に送る。

【0076】

また、類似構造判定調整手段26は、入力装置6から与えられた指定から、属性値の種別ごとに値の差異を無視するか否かを決定し、その指定項目を、頻出類似パターン検出手段24に送る。

【0077】

類似構造生成手段22は、類似構造生成調整手段25からの指定に従って、言語解析手

段 21 より与えられた集合中の各構造の部分構造について、該部分構造の類似構造の生成を行い、生成された各類似構造を、それぞれの生成元の部分構造の同値類とする。

【0078】

頻出類似パターン検出手段 24 は、類似構造判定調整手段 26 からの指定に従って、属性値の差異の無視を行いながら、類似構造生成手段 22 より与えられた同値類の集合から頻出パターンの検出を行う。

【0079】

図 7 は、本発明の第 3 の実施の形態の動作を説明するための流れ図である。次に、図 6 及び図 7 のフローチャートを参照して本発明の第 3 の実施の形態の動作について詳細に説明する。

【0080】

まず、言語解析手段 21 がテキスト DB 11 からテキスト集合を読み込む。

【0081】

言語解析手段 21 は、テキスト集合中の各テキストに対して解析を行い、解析結果として文構造を生成し、類似構造生成手段 22 に送る（図 7 のステップ A1）。図 7 のステップ A1 における言語解析手段 21 の動作は、前記第 1 の実施の形態の言語解析手段 21 と同一である。

【0082】

次に、入力装置 6 が、使用者から文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、属性値の種別ごとに値の差異を無視するか否かを指定するための入力とを受け付け、それぞれ類似構造生成調整手段 25 と類似構造判定調整手段 26 に送る（図 7 のステップ C1）。

【0083】

類似構造生成調整手段 25 は、入力装置 6 から与えられた指定から、文構造の差異の種別ごとに同一構造と判定するか否かの指定項目を生成し、類似構造生成手段 22 に送る。また、類似構造判定調整手段 26 は、入力装置 6 から与えられた指定から属性値の種別ごとに値の差異を無視するか否かの指定項目を生成し頻出類似パターン検出手段 24 に送る（図 7 のステップ C2）。

【0084】

類似構造生成手段 22 は、類似構造生成調整手段 25 からの指定に従って、言語解析手段 21 より与えられた集合中の各文構造の部分構造の類似構造の生成を行い、生成された各類似構造を、それぞれの生成元の部分構造の同値類とし、同値類の集合を頻出類似パターン検出手段 24 に送る（図 7 のステップ C3）。

【0085】

頻出類似パターン検出手段 24 は、類似構造判定調整手段 26 からの指定に従って属性値の無視を行いながら、類似構造生成手段 22 より与えられた同値類の集合から頻出パターンの検出を行う（図 7 のステップ C4）。

【0086】

最後に、頻出類似パターン検出手段 24 は、検出した頻出パターンを出力装置 3 に出力する（図 7 のステップ A4）。

【0087】

図 8 は、図 7 のステップ C3 における、類似構造生成手段 22 の動作の詳細なフローチャートである。

【0088】

図 8 を参照すると、類似構造生成手段 22 は、

ステップ C3-1 の判定において、並列の変形が指定されている場合、並列の変形（図 8 のステップ A2-1）を行い、部分構造の生成（図 8 のステップ A2-2）を行い、並列の変形が指定されていない場合、ステップ A2-2 へ分岐する。並列の変形、部分構造の生成は、図 3 のステップ A2-1、A2-2 と同一である。

【0089】

ステップC3-2の判定において、有向枝の無向枝化が指定されている場合、有向枝の無向枝化(図8のステップA2-3)を行い、指定されていない場合、ステップC3-3に分岐する。有向枝の無向枝化は、図3のステップA2-3と同一である。

【0090】

ステップC3-3の判定において、同義語の置換が指定されている場合、同義語の置換(図8のステップA2-4)を行い、同義語の置換が指定されていない場合、ステップC3-4に進む。同義語の置換は、図3のステップA2-4と同一である。

【0091】

ステップC3-3の判定において、順序木の無順序木化が指定されている場合、順序木の無順序木化(図8のステップA2-5)を行い、指定されていない場合、ステップA2-6に分岐する。

【0092】

ステップA2-6では、同値類を生成する。順序木の無順序木化、同値類を生成は、図3のステップA2-5、A2-6と同一である。

【0093】

このように、本発明の第3の実施の形態では、並列の変形(図8のステップA2-1)、有向枝の無向枝化(図8のステップA2-3)、同義語の置換(図8のステップA2-4)、及び、順序木の無順序木化(図8のステップA2-5)が、類似構造生成調整手段25から与えられた指定により、実行の有無が制御される点で、図3に示した前記第1の実施の形態の類似構造生成手段22と相違している。

【0094】

使用者は、出力されたパターンを参照して、ステップC1に戻りどこまで類似した構造を同一と判定するかを指定するための入力を再度行ったうえで本発明に頻出パターン検出を再度行わせることができる。

【0095】

次に、本発明の第3の実施の形態の作用効果について説明する。

【0096】

本発明の第3の実施の形態では、類似構造生成調整手段と類似構造判定調整手段が使用者からの指定に基づきどこまで類似した構造を同一な構造と判定するかの調整を行うように構成されているため、使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整できる。

【0097】

次に、本発明の第4の実施の形態について図面を参照して詳細に説明する。

【0098】

図9を参照すると、本発明の第4の実施の形態は、前記した第1、第2、第3の実施の形態をプログラムにより構成した場合に、そのプログラムにより動作されるコンピュータの構成を示す図である。

【0099】

テキストマイニング用プログラム7は、データ処理装置8に読み込まれ、データ処理装置8の動作を制御する。データ処理装置8はテキストマイニング用プログラム7の制御により以下の処理、すなわち第1、第2及び第3の実施の形態におけるデータ処理装置2、4及び5による処理と同一の処理を実行する。

【実施例】

【0100】

次に、本発明を具体的な実施例を即して詳細に説明する。

【0101】

まず、本発明の第1の実施例を、図面を参照して説明する。本発明の第1の実施例は、前記第1の実施の形態の一具体例である。

【0102】

本発明の第1の実施例においては、図1のデータ処理装置2をパーソナル・コンピュー

タで構成し、記憶装置 1 を、磁気ディスク記憶装置で構成し、出力装置 3 としてディスプレイを備えている。

【0103】

パーソナル・コンピュータ 2 は、言語解析手段 21、類似構造生成手段 22、頻出パターン検出手段 23 として機能する中央演算装置 (CPU) を有している。磁気ディスク記憶装置には、テキスト DB 11 としてテキスト集合が記憶されている。

【0104】

図 15 は、テキスト集合の内容を示す図である。

【0105】

言語解析手段 21 は、テキスト DB 11 中の図 15 に示されるテキスト集合の各テキストに対して言語解析を行い、各テキストの文構造を得る (図 2 のステップ A1)。

【0106】

図 16 (a) 乃至 (c) に、言語解析手段 21 で得られる文構造 (文 1 ~ 文 3) を示す。

【0107】

次に、類似構造生成手段 22 は、図 16 (a) 乃至 (c) に示される各文構造の部分構造の全ての類似構造を生成し、生成された類似構造を、生成元の部分構造の同値類とする (図 2 のステップ A2)。

【0108】

本実施例では、図 16 (b) に示される文 2 (「速く安い車種 A」) の文構造から、部分構造の同値類を生成する様子を例にとって説明する。この例は、図 18 乃至 21 に示されている。

【0109】

まず、図 18 に示すように、並列構造の変形を行う (図 3 のステップ A2-1)。図 18 に示す例においては、部分構造 2a-0 において、並列関係にある「速い」と「安い」の接続関係を変形し、類似構造 2a-1 を生成している。

【0110】

次に、図 19 に示すように、部分構造の生成を行う (図 3 のステップ A2-2)。図 19 に示す例においては、部分構造 2a-0 から、2 単語の関係を表す部分構造 2c-0 及び 2g-0 と、1 単語の部分構造 2d-0、2e-0 及び 2f-0 を生成する。

【0111】

また、類似構造 2a-1 から、部分構造 2a-0 に含まれない 2 単語の関係を表す部分構造 2b-0 を生成する。

【0112】

なお、部分構造 2a-0 と類似構造 2a-1 の両方から生成される構造は 1 つにまとめて扱う。

【0113】

また、ここで部分構造を生成するのに用いた部分構造 2a-0、及び類似構造 2a-1 も、今後の類似構造生成において、部分構造及び類似構造として扱う。

【0114】

次に、図 20 に示すように、有向枝の無向枝化を行う (図 3 のステップ A2-3)。この例においては、ステップ A2-2 において生成した部分構造の全ての有向枝を無向枝化して、新たな類似構造を生成している。図 20 (a) に示すように、例えば部分構造 2a-0 の有向枝を無向枝化して、類似構造 2a-2 が生成される。なお、1 単語からなり有向枝を持たない部分構造 2d-0、2e-0 及び 2f-0 は、ステップ A2-3 では変形が行われないため、図 20 では省略されている。

【0115】

次に、同義語の置換が行われる (図 3 のステップ A2-4)。本実施例における「同義語の置換」では、ユーザによりあらかじめ与えられた同義語辞書に定義された被置換語を代表語に置き換えるものとする。

【0116】

また、本実施例に用いる同義語辞書は、図17に示されるように、被置換語「高速」を代表語「速い」に置き換える1つの辞書項目のみが登録された同義語辞書が指定されたものとしている。

【0117】

ここで、図20に示された、この時点で生成された部分構造及び類似構造には、被置換語「高速」が含まれないため、ステップA2-4では、変形が発生しない。そのため、ここではステップA2-4による変形の図を省略している。

【0118】

次に、順序木の無順序木化が行われる（図3のステップA2-5）。ここでは、文構造の木構造において、兄弟関係にある単語を50音順にソートすることによって、順序木の無順序木化を行う。

【0119】

なお、順序木の無順序木化を行うための他の方法として、

- ・兄弟関係にある単語を50音順以外の一定の法則に従いソートする方法や、
- ・ソートを行わずに頻出類似パターン検出時に兄弟関係にある単語の順序だけが異なる木を同一と判定する方法を用いてもよい。

【0120】

図20に示された、この時点で、生成された部分構造及び類似構造では、類似構造2a-1及び2a-3を除いた部分構造及び類似構造では兄弟関係になっている単語が存在せず、類似構造2a-1及び2a-3では、既に兄弟関係にある単語が50音順に並んでいるため、実質的に変形が発生しない。そのため、ここではステップA2-5による変形の図を省略している。

【0121】

最後に、類似構造を生成元の部分構造の同値類とすることで、同値類の生成を行う（図3のステップA2-6）。

【0122】

図20に示された部分構造及び類似構造の集合において、各類似構造を生成元の部分構造の同値類することで生成される同値類を図21に示す。部分構造2a-0と、部分構造2a-0の有向枝を無向枝化することで生成された類似構造2a-2と、部分構造2a-0を並列変形した類似構造2a-1と、類似構造2a-1の有向枝を無向枝化することで生成された類似構造2a-3とは同値類2aを構成している。部分構造2b-0と、部分構造2b-0の有向枝を無向枝化することで生成された類似構造2b-1は、同値類2bを構成している。部分構造2c-0と、部分構造2c-0の有向枝を無向枝化することで生成された類似構造2c-1は、同値類2cを構成している。部分構造2g-0と、部分構造2g-0の有向枝を無向枝化することで生成された類似構造2g-1は、同値類2gを構成している。部分構造2d-0、2e-0、2f-0は、類似構造と部分構造は同一である。

【0123】

図18乃至図21に示したように、本実施例において、文2の文構造から、類似構造生成手段22が同値類を生成する例においては、同義語の置換（図3のステップA2-4）及び順序木の無順序木化（図3のステップA2-5）で変形は行われない。

【0124】

図22に示すように、文3の文構造の1つの部分構造に対して、類似構造生成手段22が行う変形を用いて、同義語の置換（図3のステップA2-4）及び順序木の無順序木化（図3のステップA2-5）で発生する変形の例を説明する。

【0125】

まず文3の文構造を表す部分構造3a-0に対して並列の変形（図3のステップA2-1）が行われる。ここでは、部分構造3a-0が並列の構造を含まず変形が行われなかったため、図22には、並列の変形による結果の構造は含まれない。

【0126】

次に、部分構造 3a-0 から部分構造の生成（図 3 のステップ A2-2）が行われる。ここでは、部分構造 3a-0 に行われる構造変形にのみ注目して説明するため、部分構造 3a-0 から、他の部分構造を生成する処理である部分構造の生成は省略する。

【0127】

次に、部分構造 3a-0 に対して、有向枝の無向枝化（図 3 のステップ A2-3）が行われる。図 22（a）の部分構造 3a-0 の「安い」から、「車種 A」への有向枝と「高速」から「車種 A」への有向枝が無向枝化され、図 22（b）の類似構造 3a-1 が生成される。

【0128】

次に、類似構造 3a-1 に対して、同義語の置換（図 3 のステップ A2-4）が行われる。ここでは、図 17 に示される同義語辞書を用いているため、被置換語「高速」が代表語「速い」に置き換えられる。類似構造 3a-1 に含まれる被置換語「高速」も代表語「速い」に置き換えられ、図 22（c）の類似構造に変形される。

【0129】

次に、図 22（c）の類似構造 3a-1 に対して、順序木の無順序木化（図 3 のステップ A2-5）が行われる。ここでは、兄弟関係にある単語を、50 音順にソートすることで順序木の無順序木化が行われるため、類似構造 3a-1 において兄弟関係にある「安い」と「速い」の順序を入れ替え、50 音順にソートにされ、図 22（d）の類似構造に変換される。

【0130】

このようにして生成された類似構造に対して同値類の生成（図 3 のステップ A2-6）が行われるが、本実施例では、部分構造 3a-0 から生成される一つの類似構造 3a-1 に行われる変形のみ注目して説明しているため省略する。

【0131】

このようにして、類似構造生成手段 22 が、部分構造と類似構造及び同値類の生成を行うことで、本実施例では、図 16 の文 1 の文構造から、図 23 に示すような同値類が生成される。図 16 の文 2 の文構造から、図 24 に示すような同値類が生成される。また図 16 の文 3 の文構造から、図 25 に示すような同値類が生成される。

【0132】

ただし、本来は、図 22 における変形の途中経過（図 22（b）、図 22（c）の類似構造 3a-1）のように、形の違う類似構造も生成されているが、説明を分かりやすくするため、頻出パターンの検出に用いられない構造は、図 23 乃至図 25 の同値類からは省略している。

【0133】

次に、頻出パターン検出手段 23 は、図 23 乃至図 25 に示される同値類の集合から頻出パターン（頻出する同値類）の検出を行う（図 2 のステップ A3）。

【0134】

この際、頻出パターン検出手段 23 は、要素の少なくとも一つが同一である同値類は、同一と判定して、頻出パターンの検出を行う。

【0135】

例えば、本実施例においては、図 23 の同値類 1c の要素である類似構造 1c-1 と、図 24 の同値類 2b の要素である類似構造 2b-1 は、どちらも「車種 A」と「速い」が無向枝で連結された構造で、属性値の差分もないため、同一の構造である。

【0136】

従って、頻出パターン検出手段 23 は、図 23 の同値類 1c と図 24 の同値類 2b を同一と判定する。

【0137】

図 23 乃至図 25 を参照すると、

「類似構造 1c-1、類似構造 2b-1 と、類似構造 3c-1」、

「部分構造 1 d - 0、部分構造 2 d - 0 と、類似構造 3 e - 1」、
「部分構造 1 e - 0、部分構造 2 f - 0 と、部分構造 3 f - 0」、
「部分構造 1 f - 0 と部分構造 2 e - 0」
がそれぞれ同一の構造となっている。

【0138】

「要素の少なくとも一つが同一である同値類は同一と判定する」という同値類の性質により、図 2 3 乃至図 2 5 に示される同値類のうち、

「同値類 1 c、2 b、及び、3 c」、
「同値類 1 d、2 d、及び、3 e」、
「同値類 1 e、2 f、及び、3 f」、
「同値類 1 f、及び、2 e」

がそれぞれ同一の同値類と判定される。

【0139】

本実施例では、3 回以上出現する同値類を頻出パターンとする。なお、どのような出現回数の同値類を頻出パターンとして検出するかは、使用者がテキストマイニングを実行する前に決定することができる。

【0140】

この場合、

「同値類 1 c、2 b、及び、3 c」、
「同値類 1 d、2 d、及び、3 e」、
「同値類 1 e、2 f、及び、3 f」

が頻出パターンとして検出される。

【0141】

最後に、そのようにして抽出された頻出パターンを表す構造を出力装置 3 に表示する（図 2 のステップ A 4）。

【0142】

図 2 6 は、本実施例において、出力装置 3 が出力する頻出パターンの表現の一例を示す図である。本実施例では、頻出パターンを表す同値類の要素である類似構造を、頻出パターンの表現として用いている。

【0143】

このようにして、類似構造を生成し、同値類を生成して頻出パターンの検出を行うことで、「部分構造 1 c - 0（図 2 3）、部分構造 2 b - 0（図 2 4）、及び、部分構造 3 c - 0（図 2 5）」のように類似した意味を持つが、連結構造の異なる部分構造を同一と判定し、頻出パターンとして検出することができる。

【0144】

次に、本発明の第 2 の実施例を、図面を参照して説明する。本実施例は、前記第 2 の実施の形態に対応するものである。

【0145】

本発明の第 2 の実施例は、データ処理装置 4 をパーソナル・コンピュータで構成し、記憶装置 1 を磁気ディスク記憶装置で構成し、出力装置 3 としてディスプレイを備えている。

【0146】

パーソナル・コンピュータ 4 は、言語解析手段 2 1、類似構造生成手段 2 2、頻出類似パターン検出手段 2 4 として機能する中央演算装置（CPU）を有し、磁気ディスク記憶装置には、テキスト DB 1 1 としてテキスト集合が記憶されている。テキスト集合としては、前記第 1 の実施例と同様、図 1 5 に示した文 1 乃至文 3 を使用する。

【0147】

言語解析手段 2 1 は、テキスト DB 1 1 中の図 1 5 に示されるテキスト集合の各テキストに対して、言語解析を行い、各テキストの文構造を得る（図 5 のステップ A 1）。ここで得られる文構造は、前記第 1 の実施例と同様、図 1 6 のようになる。

【0148】

次に、類似構造生成手段 22 は、図 16 に示される各文構造の部分構造の全ての類似構造を生成し、生成された類似構造を生成元の部分構造の同値類とする（図 5 のステップ A 2）。ここで得られる同値類は、前記第 1 の実施例と同様、図 23 乃至図 25 のようになる。

【0149】

次に、頻出類似パターン検出手段 24 は、図 23 乃至図 25 に示される同値類の集合から、属性値の差異を無視しながら頻出パターン（頻出する同値類）の検出を行う（図 5 のステップ B 3）。

【0150】

頻出類似パターン検出手段 24 は、要素の少なくとも一つが同一である同値類は同一と判定して、頻出パターンの検出を行う。ただし、本実施例の頻出類似パターン検出手段 24 は、表層格や付属語情報などの属性値の差異を無視して、類似構造の同一性の判定を行っており、この点で、前記第 1 の実施例の頻出パターン検出手段 23 と相違している。

【0151】

例えば、図 23 の類似構造 1a-1 と図 24 の類似構造 2a-3 は、どちらも、「車種 A」と「速い」及び「安い」が無向枝で連結された構造であるが、表層格が異なるため、前記第 1 の実施例の頻出パターン検出手段 23 では、同一と判定されない。一方、本実施例の頻出類似パターン検出手段 24 では、同一と判定される。

【0152】

本実施例においては、図 23 乃至図 25 を参照すると、

「類似構造 1a-1、類似構造 2a-3、及び、類似構造 3a-1」、
「類似構造 1b-1、類似構造 2c-1 と、類似構造 3b-1」、
「類似構造 1c-1、類似構造 2b-1、及び、類似構造 3c-1」、
「部分構造 1d-0、部分構造 2d-0、及び、類似構造 3e-1」、
「部分構造 1e-0、部分構造 2f-0、及び、部分構造 3f-0」、
「部分構造 1f-0、部分構造 2e-0、及び、部分構造 3d-0」

がそれぞれ頻出類似パターン検出手段 24 に同一の構造と判定される。

【0153】

頻出類似パターン検出手段 24 は、要素の少なくとも一つが同一である同値類は同一と判定するため、

「同値類 1a、2a、及び、3a」、
「同値類 1b、2c、及び、3b」、
「同値類 1c、2b、及び、3c」、
「同値類 1d、2d、及び、3e」、
「同値類 1e、2f、及び、3f」、
「同値類 1f、2e、及び、3d」

をそれぞれ同一の同値類と判定する。

【0154】

本実施例では、前記第 1 の実施例と同様に、3 回以上出現する同値類を頻出パターンとする。

【0155】

この場合、

「同値類 1a、2a、及び、3a」、
「同値類 1b、2c、及び、3b」、
「同値類 1c、2b、及び、3c」、
「同値類 1d、2d、及び、3e」、
「同値類 1e、2f、及び、3f」、
「同値類 1f、2e、及び、3d」

が頻出パターンとして検出される。

【0156】

最後に、そのようにして抽出された頻出パターンを表す構造を、出力装置3に表示する(図5のステップA4)。

【0157】

本実施例において、出力装置3が出力する頻出パターンの表現は図27のようになる。本実施例では、前記第1の実施例と同様に、頻出パターンを表す同値類の要素である類似構造を、頻出パターンの表現として用いている。

【0158】

このようにして、属性値の差異を無視して頻出パターンの検出を行うことで、

「部分構造1b-0(図23)、部分構造2c-0(図24)と部分構造3b-0(図25)」、

「部分構造1f-0(図23)、部分構造2e-0(図24)と部分構造3f-0(図25)」

のように、類似した意味を持つが属性値の異なる部分構造を同一と判定し、頻出パターンとして、検出を行うことができる。

【0159】

次に、本発明の第3の実施例を、図面を参照して説明する。本実施例は、本発明の第3の実施の形態に対応するものである。

【0160】

本発明の第3の実施例は、データ処理装置5をパーソナル・コンピュータで構成し、記憶装置1を磁気ディスク記憶装置で構成し、出力装置3としてディスプレイを、入力装置6としてキーボードを備えている。

【0161】

パーソナル・コンピュータ5は、言語解析手段21、類似構造生成手段22、頻出類似パターン検出手段24、類似構造生成調整手段25、類似構造判定調整手段26として機能する中央演算装置(CPU)を有している。磁気ディスク記憶装置には、テキストDB11としてテキスト集合が記憶されている。テキスト集合としては、前記第1、第2の実施例と同様、図15に示した文が用いられる。

【0162】

言語解析装置21は、テキストDB11中の図15に示されるテキスト集合の各テキストに対して、言語解析を行い、各テキストの文構造を得る(図7のステップA1)。ここで得られる文構造は、前記第1、第2の実施例と同じく、図16のようになる。

【0163】

次に、使用者は、入力装置6を用いて、

- ・文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、
 - ・属性値の種別ごとに値の差異を無視するか否かを指定するための入力
- を行う(図7のステップC1)。

【0164】

本実施例では、

「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異は同一と判定する」という入力を行ったとする。

【0165】

入力装置6は、使用者から受け付けた入力を、類似構造生成調整手段25と類似構造判定調整手段26に送る。

【0166】

次に、類似構造生成調整手段25は、入力装置6から使用者の指定を受け取り、類似構造生成手段22の動作を制御する(図7のステップC2)。

【0167】

本実施例においては、類似構造生成調整手段25は、入力装置6から、

「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異は同一と判定する」

という指定を受け取り、類似構造生成手段 22 が、文構造の部分構造から類似構造を生成する際の変形処理において、並列構造の変形（図 8 のステップ A 2-1）、有向枝の無向枝化（図 8 のステップ A 2-3）及び順序木の無順序木化（図 8 のステップ A 2-5）は行われるが、同義語の置換（図 8 のステップ A 2-4）はスキップされるように制御する。

【0168】

一方、類似構造判定調整手段 26 は、入力装置 6 から使用者の入力を受け取り、頻出類似パターン検出手段 24 の動作を制御する（図 7 のステップ C 2）。

【0169】

本実施例においては、類似構造生成調整手段 26 は、入力装置 6 から、「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異については同一と判定する」という指定を受け取り、頻出類似パターン検出手段 24 が類似構造の同一性判定の処理を、表層格の差異を無視し、付属語情報の差異も無視して行うように制御する。

【0170】

次に、類似構造生成手段 22 は、図 16 に示される各文構造の部分構造についてステップ C 2 で生成した指定項目に従い、同義語の置換（図 8 のステップ A 2-4）を飛ばして類似構造を生成し、生成された類似構造を生成元の部分構造の同値類とする（図 7 のステップ C 3）。

【0171】

以下、図 16 に示される文 3 の文構造の 1 つの部分構造に対して、類似構造生成手段 22 が行う変形を例にとって説明する。図 28 に、一例を示す。

【0172】

まず、文 3 の文構造を表す部分構造 3a-0 に対して、並列の変形（図 8 のステップ A 2-1）が行われる。ただし、図 28 に示す例では、部分構造 3a-0 が並列の構造を含まず変形が行われないため、図 28 には、並列の変形による結果の構造は含まれない。

【0173】

次に、部分構造 3a-0 から部分構造の生成（図 8 のステップ A 2-2）が行われるが、ここでは、部分構造 3a-0 に行われる構造変形にのみ注目して説明するため、部分構造 3a-0 から他の部分構造を生成する処理である部分構造の生成は省略する。

【0174】

次に、部分構造 3a-0 に対して有向枝の無向枝化（図 8 のステップ A 2-3）が行われる。部分構造 3a-0 の「安い」から「車種 A」への有向枝と、「高速」から「車種 A」への有向枝が無向枝化され、図 28（b）の類似構造 3a-2 が生成される。

【0175】

同義語の置換（図 8 のステップ A 2-4）は、類似構造生成調整手段 25 より与えられた指定により、ステップ C 3-3 の判定でスキップされるため、実行されない。

【0176】

次に、類似構造 3a-2 に対して、順序木の無順序木化（図 8 のステップ A 2-5）が行われる。ここでは、兄弟関係にある単語を 50 音順にソートすることで、順序木の無順序木化が行われる。図 28（b）の類似構造 3a-2 において、兄弟関係にある「安い」と「高速」の順序を入れ替えて、50 音順にソートされ、図 28（c）の構造に変換される。

【0177】

このようにして生成された類似構造に対して、同値類の生成（図 8 のステップ A 2-6）が行われるが、ここでは、部分構造 3a-0 から生成される一つの類似構造 3a-2 に

行われる変形のみ注目して説明しているため、省略する。

【0178】

本実施例における変形では、同義語の置換（図8のステップA2-4）が飛ばされるため、図28（c）の類似構造3a-2には、被置換語「高速」が残っている。一方、図22に示した前記第1、第2の実施例における変形の例では、図22（d）の類似構造3a-1では、被置換語「高速」が代表語「速い」に置換されている。

【0179】

本実施例では、このようにして、類似構造生成手段22が部分構造と類似構造及び同値類の生成を行うことで、図16の文1の文構造から、図23に示される同値類が生成され、図16の文2の文構造から、図24に示される同値類が生成され、図16の文3の文構造から図29に示される同値類が生成される。

【0180】

次に、頻出類似パターン検出手段24は、図23、図24、及び図29に示される同値類の集合から、ステップC2で、類似構造判定調整手段26が指定した属性値の差異を無視しながら頻出パターンの検出を行う（図7のステップC4）。

【0181】

頻出類似パターン検出手段24は、要素の少なくとも一つが同一である同値類は同一と判定して、頻出パターンの検出を行う。

【0182】

本実施例においては、頻出類似パターン検出手段24は、類似構造判定調整手段26からの指定により、どの属性値の差異を無視して類似構造の同一性を判定するかを決定する。

【0183】

本実施例では、

「表層格の差異を無視する」、

「付属語情報の差異を無視する」

と動作を制御するように類似構造判定調整手段26が指定を行ったため、頻出類似パターン検出手段24は、前記第2の実施例と同様に、類似構造の同一性の判定を行う。

【0184】

本実施例においては、図23、図24、及び図29を参照すると、

「類似構造1a-1、及び、類似構造2a-3」、

「部分構造2c-0、及び、部分構造3b-0」、

「類似構造1b-1、類似構造2c-1、及び、類似構造3b-1」、

「部分構造1c-0、及び、類似構造2b-0」、

「類似構造1c-1、及び、類似構造2b-1」、

「部分構造1d-0、及び、部分構造2d-0」、

「部分構造1e-0、部分構造2f-0、及び、部分構造3f-0」、

「部分構造1f-0、部分構造2e-0、及び、部分構造3d-0」

がそれぞれ頻出類似パターン検出手段24に同一の構造と判定される。

【0185】

頻出類似パターン検出手段24は、要素の少なくとも一つが同一である同値類は同一と判定するため、

「同値類1a、及び、2a」、

「同値類1b、2c、及び、3b」、

「同値類1c、及び、2b」、

「同値類1d、及び、2d」、

「同値類1e、2f、及び、3f」、

「同値類1f、2e、及び、3d」

をそれぞれ同一の同値類と判定する。

【0186】

本実施例では、前記第1、第2の実施例と同様に、3回以上出現する同値類を頻出パターンとする。

【0187】

この場合、

「同値類1b、2c、及び、3b」、

「同値類1e、2f、及び、3f」、

「同値類1f、2e、及び、3d」

が頻出パターンとして検出される。

【0188】

最後に、このようにして抽出された頻出パターンを表す構造を、出力装置3に表示する(図7のステップA4)。

【0189】

本実施例において、出力装置3が出力する頻出パターンの表現は、図30のようになる。図30に示すように、本実施例では、前記第1、第2の実施例と同様に、頻出パターンを表す同値類の要素である類似構造を頻出パターンの表現として用いている。

【0190】

使用者は、この頻出パターン検出に不満を感じた場合、図7のステップC1に戻り、どこまで類似した構造を同一と判定するか指定の入力を変更することで、再度頻出パターンの検出を行うことができる。

【0191】

このようにして、

「同義語の置換による差異については同一と判定しない」

という使用者の指定に基づき、図23、図24、図29において、

「部分構造1a-0、部分構造2a-0、及び、部分構造3a-0」、

「部分構造1c-0、部分構造2b-0、及び、部分構造3c-0」、

「部分構造1d-0、部分構造2d-0、及び、部分構造3e-0」

といった類似した意味を持つが使用者の入力に反する構造を同一と判定せずに、頻出パターン検出行うことで、使用者がどこまで類似した構造を同一と判定するか調整を行うことができる。

【産業上の利用可能性】

【0192】

本発明によれば、コンピュータ上に蓄積される、顧客からの苦情メールやアンケート結果の特徴分析を行う目的に良く用いられるテキストマイニング装置や、テキストマイニング装置をコンピュータに実現するためのプログラムといった用途に適用できる。

【図面の簡単な説明】

【0193】

【図1】本発明の第1の実施の形態の構成を示す図である。

【図2】第1の実施の形態の動作を説明するための流れ図である。

【図3】本発明の実施の形態における類似構造生成手段22の動作を説明するための流れ図である。

【図4】本発明の第2の実施の形態の構成を示す図である。

【図5】本発明の第2の実施の形態の動作を説明するための流れ図である。

【図6】本発明の第3の実施の形態の構成を示す図である。

【図7】本発明の第3の実施の形態の動作を説明するための流れ図である。

【図8】本発明の第3の実施の形態における類似構造生成手段22の動作を説明するための流れ図である。

【図9】本発明の第4の実施の形態の構成を示す図である。

【図10】従来の技術の構成を示す図である。

【図11】属性値を用いる形式で表された「彼は私が本を買ったのを知らない」という文の構文構造の例を示す図である。

【図 1 2】属性値を用いない形式で表された「彼は私が本を買ったのを知らない」という文の構文構造の例を示す図である

【図 1 3】連結構造が異なり類似した意味を持つ構造の差異の例を示す図であり、(a) は、係り受けの向きの差異、(b) は、係り受けの順序の差異、(c) は、同義語の置換による差異、(d) は、並列の構文構造と意味構造の差異を示す図である。

【図 1 4】属性値が異なり類似した意味を持つ構造の差異の複数の例を示す図である。(a) は、付属語情報の差異を示す図である。(b) は、表層格の差異を示す図である。

【図 1 5】本発明の第 1 乃至第 3 実施例で使用するテキスト DB 中のテキスト集合の例を示す図である。

【図 1 6】本発明の第 1 乃至第 3 の実施例において、図 3 に示すテキスト集合中の各テキストを言語解析して得られる文構造の集合を示す図である。

【図 1 7】本発明の第 1 乃至第 3 の実施例において使用する、同義語辞書の構造を示す図である。

【図 1 8】本発明の第 1 乃至第 3 の実施例において、図 3 のステップ A 2-1 における処理を示す図である。

【図 1 9】本発明の第 1 乃至第 3 の実施例において、図 3 のステップ A 2-2 における処理を示す図である。

【図 2 0】本発明の第 1 乃至第 3 の実施例において、図 3 のステップ A 2-3 における処理を示す図である。

【図 2 1】本発明の第 1 乃至第 3 の実施例において、図 3 のステップ A 2-6 における処理を示す図である。

【図 2 2】本発明の第 1、第 2 の実施例において、類似構造生成手段 2 2 が文 3 の文構造の全体からなる部分構造 3 a-0 の類似構造を生成する処理を示す図である。

【図 2 3】本発明の第 1 乃至第 3 の実施例において文 1 の文構造から生成される部分構造の同値類を示す図である。

【図 2 4】本発明の第 1 乃至第 3 の実施例において文 2 の文構造から生成される部分構造の同値類を示す図である。

【図 2 5】本発明の第 1、第 2 の実施例において、文 3 の文構造から生成される部分構造の同値類を示す図である。

【図 2 6】本発明の第 1 の実施例において、図 2 3 乃至 2 5 に示す同値類の集合から検出される頻出パターンを示す図である。

【図 2 7】本発明の第 2 の実施例において、図 2 3 乃至 2 5 に示す同値類の集合から検出される頻出パターンを示す図である。

【図 2 8】本発明の第 3 の実施例において、類似構造生成手段 2 2 が文 3 の文構造の全体からなる部分構造 3 a-0 の類似構造を生成する処理を示す図である。

【図 2 9】本発明の第 3 の実施例において、文 3 の文構造から生成される部分構造の同値類を示す図である。

【図 3 0】本発明の第 3 の実施例において、図 2 3、2 4 及び図 2 9 に示す同値類の集合から検出される頻出パターンを示す図である。

【符号の説明】

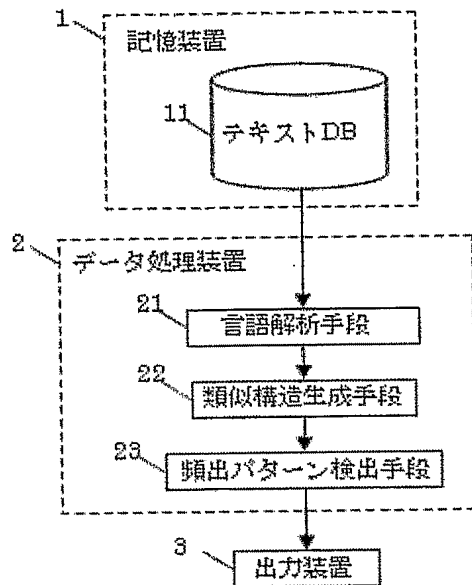
【0194】

- 1 記憶装置
- 2 データ処理装置
- 3 出力装置
- 4 データ処理装置
- 5 データ処理装置
- 6 入力装置
- 7 テキストマイニング用プログラム

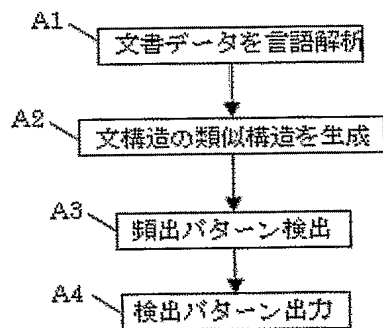
- 8 データ処理装置
- 1 1 テキストDB
- 2 1 言語解析手段
- 2 2 類似構造生成手段
- 2 3 頻出パターン検出手段
- 2 4 頻出類似パターン検出手段
- 2 5 類似構造生成調整手段
- 2 6 類似構造判定調整手段

【書類名】 図面

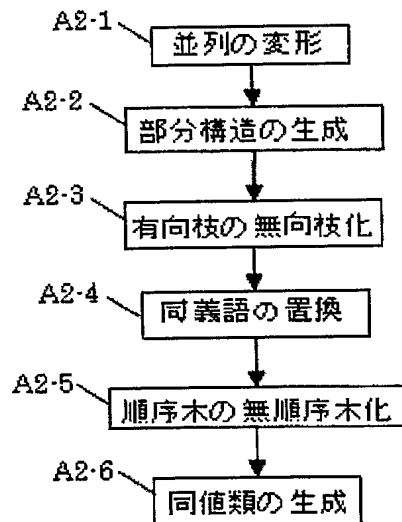
【図 1】



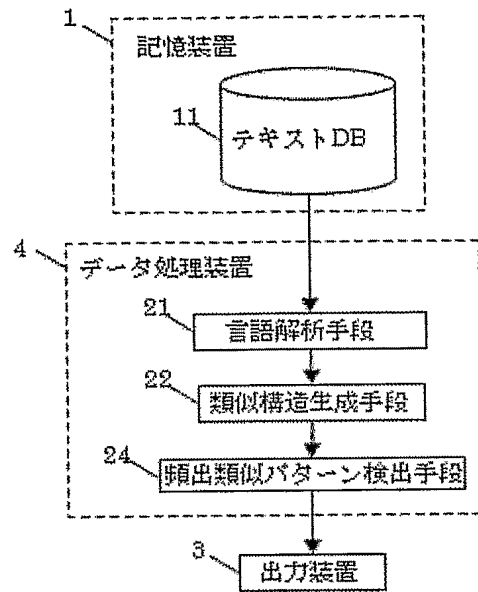
【図 2】



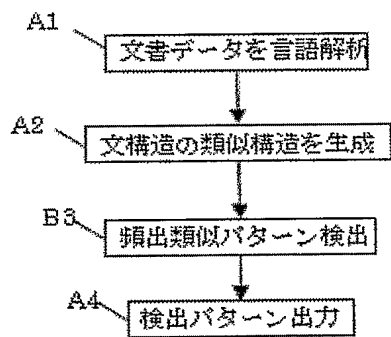
【図 3】



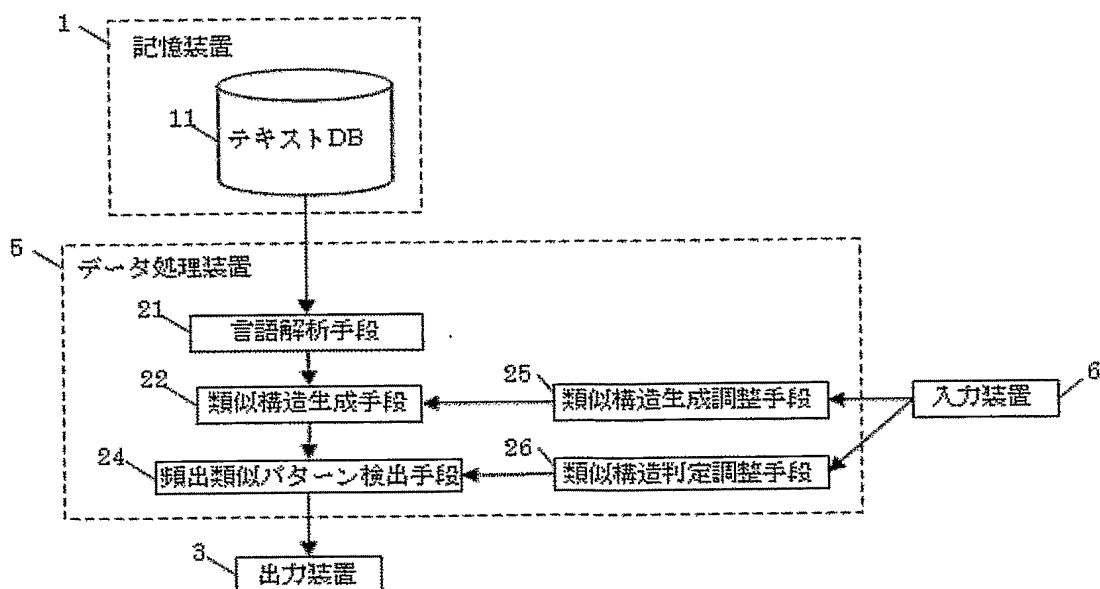
【図 4】



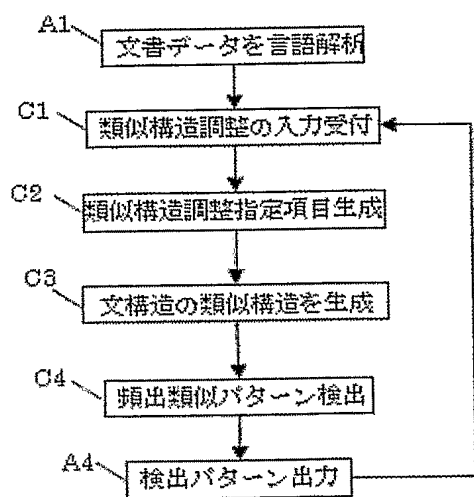
【図 5】



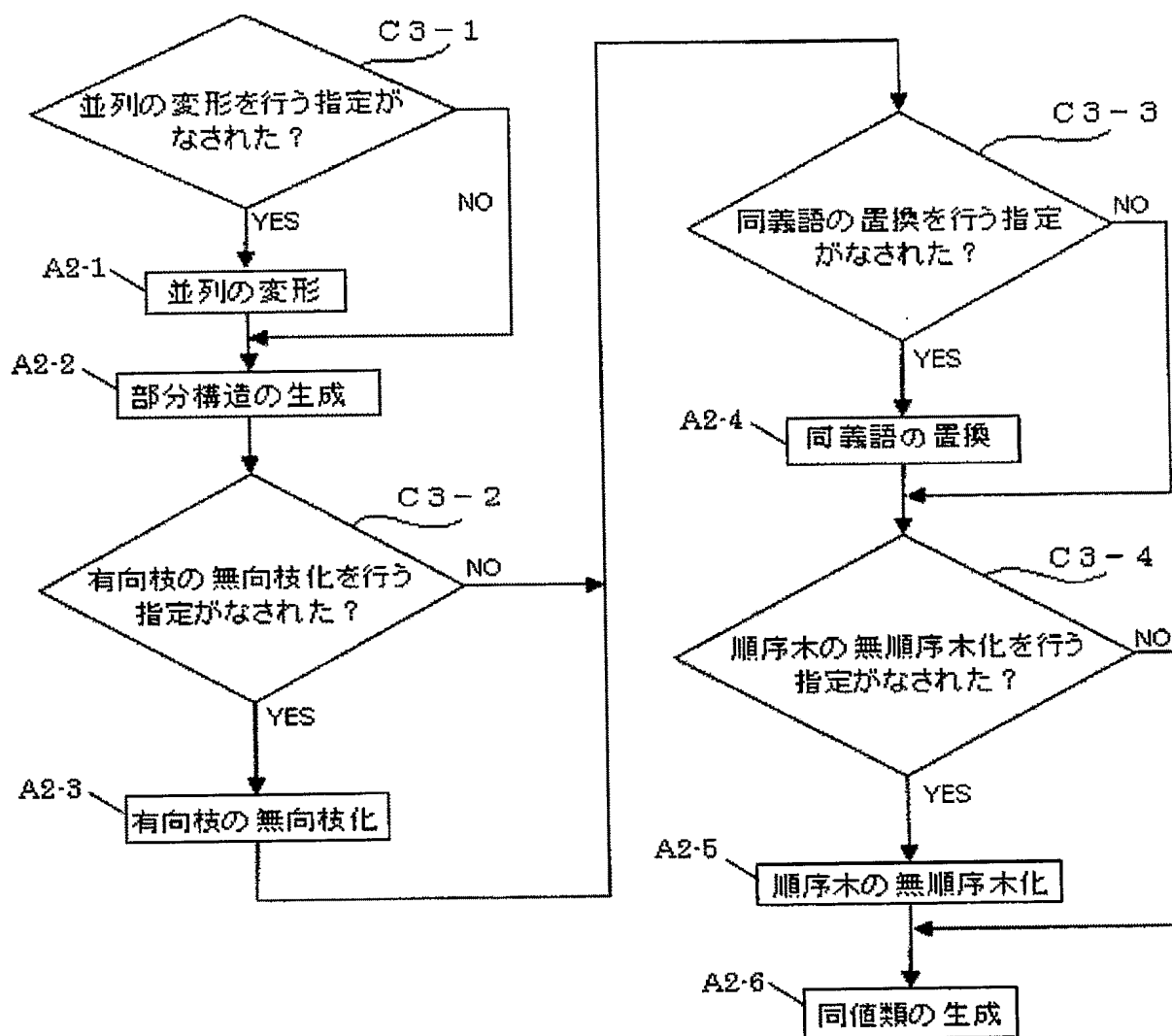
【図 6】



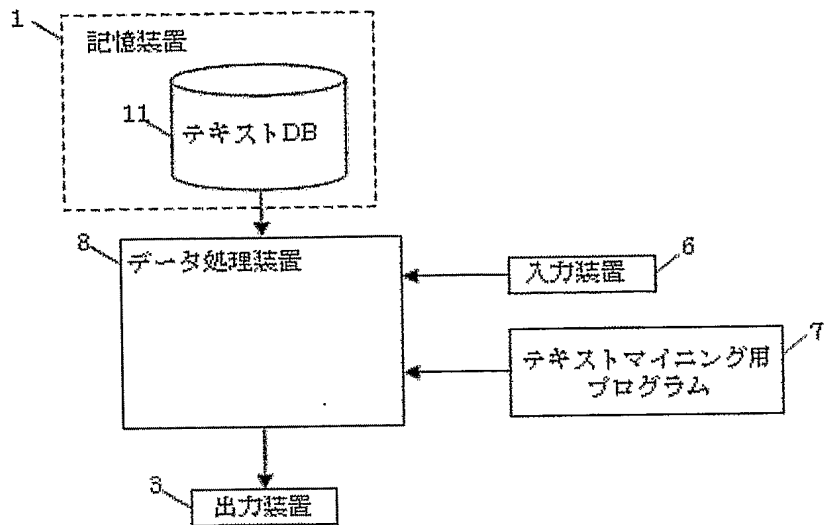
【図 7】



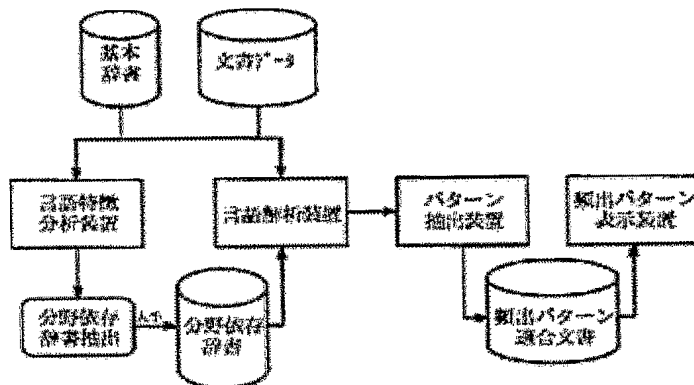
【図 8】



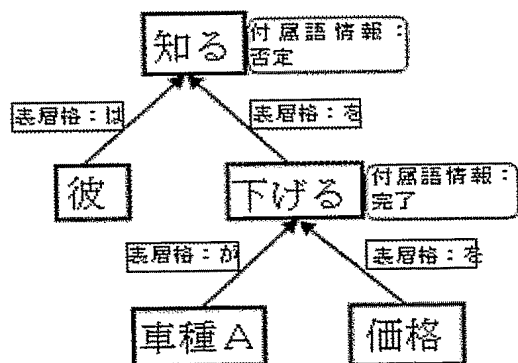
【図 9】



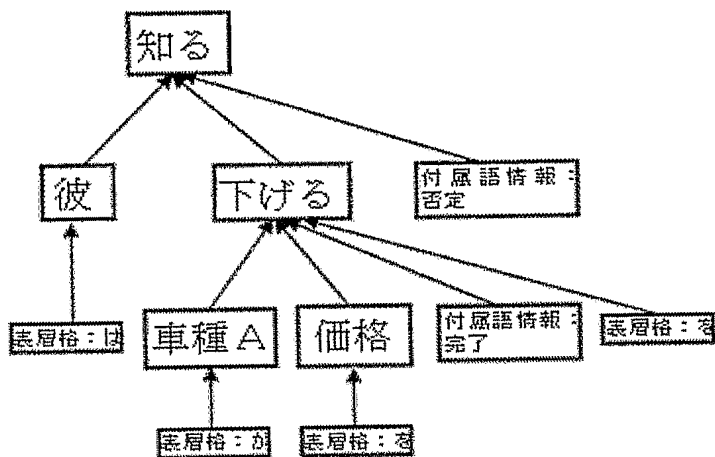
【図 10】



【図 11】

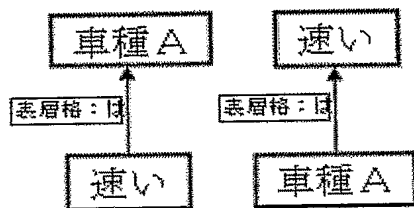


【図 12】

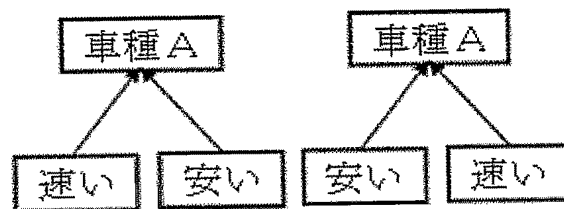


【図 13】

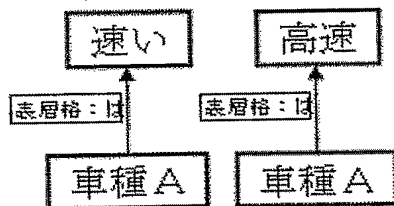
(a) 「速いのは車種 A」と
「車種 A は速い」



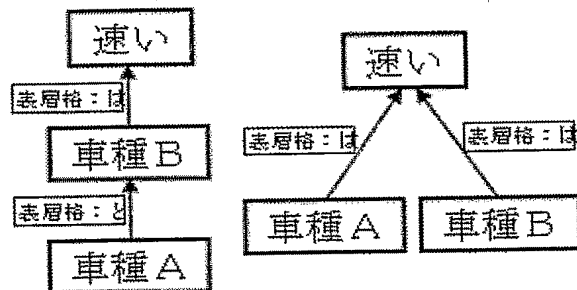
(b) 「速く安い車種 A」と「安く速い車種 A」



(c) 「車種 A は速い」と
「車種 A は高速だ」

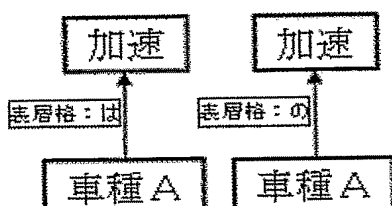


(d) 「車種 A と車種 B は速い」の
構文構造と意味構造

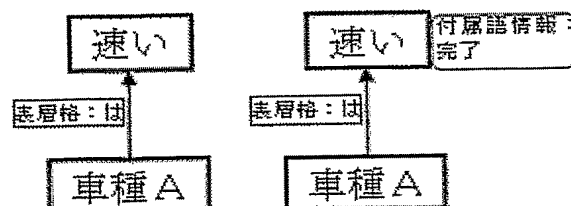


【図 14】

(a) 「車種 A は加速」と
「車種 A の加速」



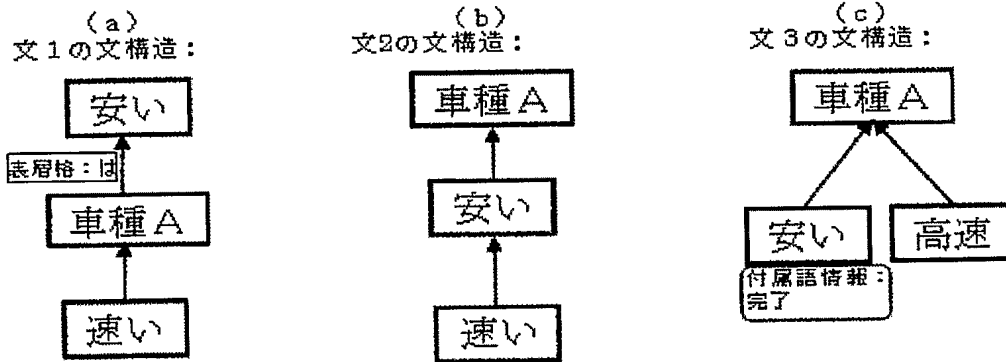
(b) 「車種 A は速い」と「車種 A は速かった」



【図 15】

文 1 : 速い車種 A は安い
 文 2 : 速く安い車種 A
 文 3 : 安かった高速な車種 A

【図 16】



【図 17】

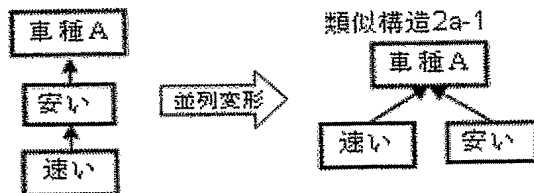
同義語辞書

代表語	被置換後
速い	高速

【図 18】

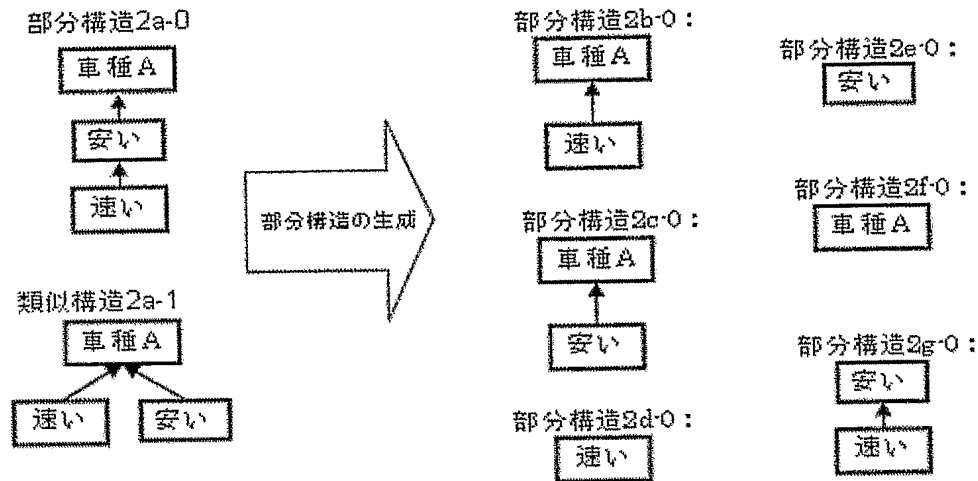
ステップ A2-1

部分構造 2a-0



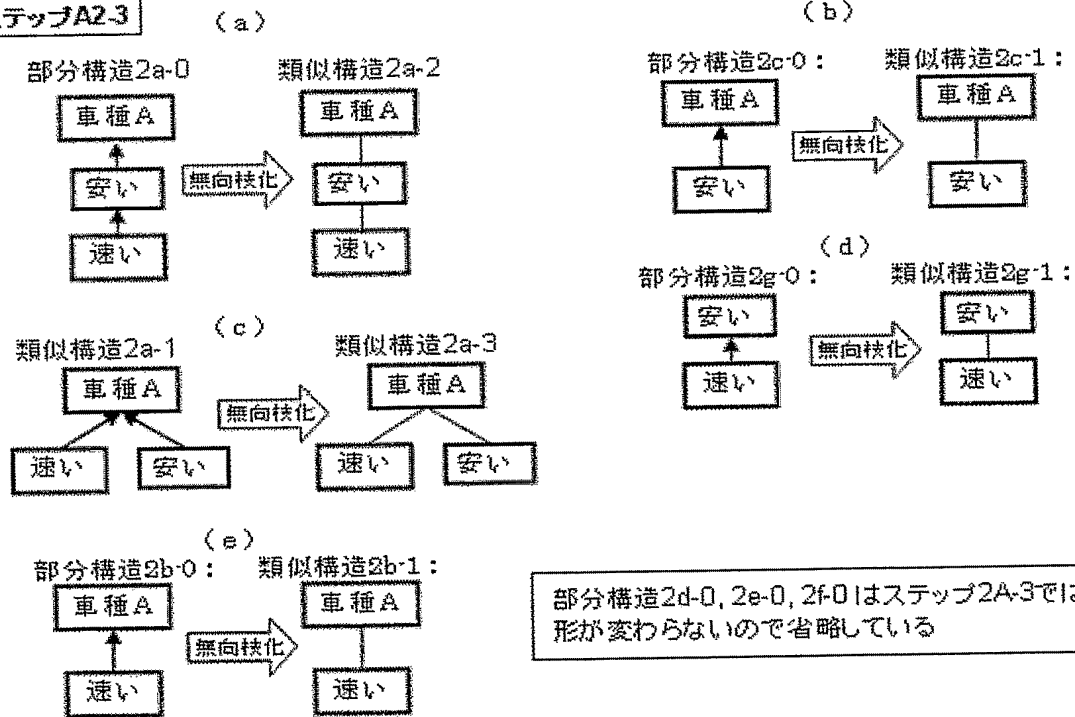
【図 19】

ステップA2-2



【図 20】

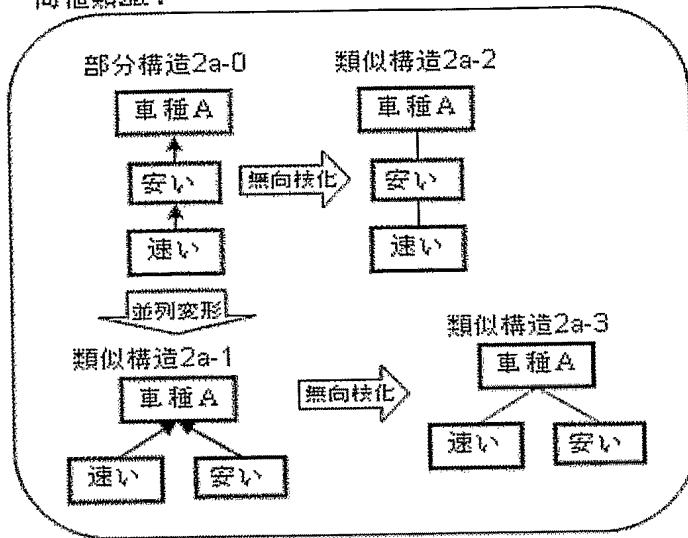
ステップA2-3



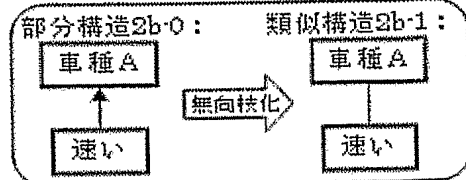
【図 2 1】

ステップA2-6

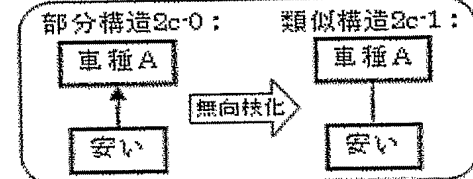
同値類2a:



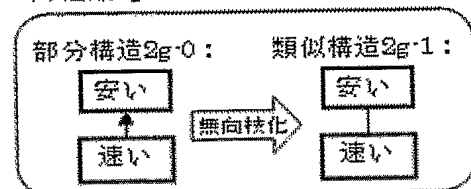
同値類2b:



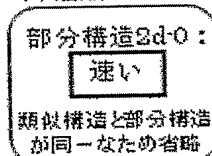
同値類2c:



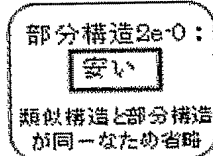
同値類2g:



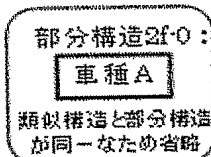
同値類2d:



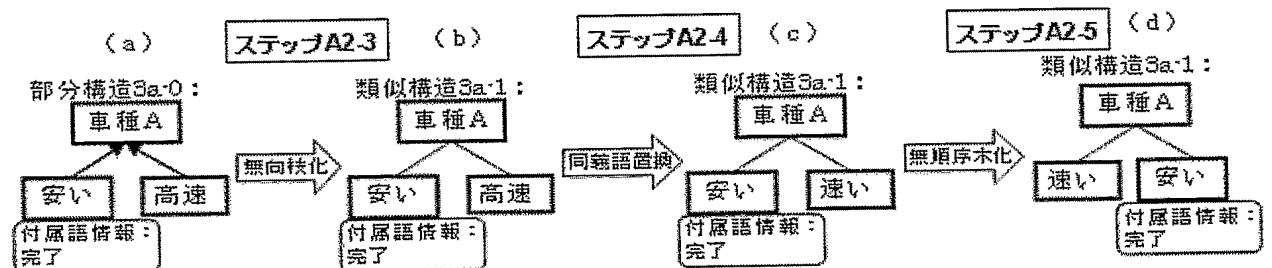
同値類2e:



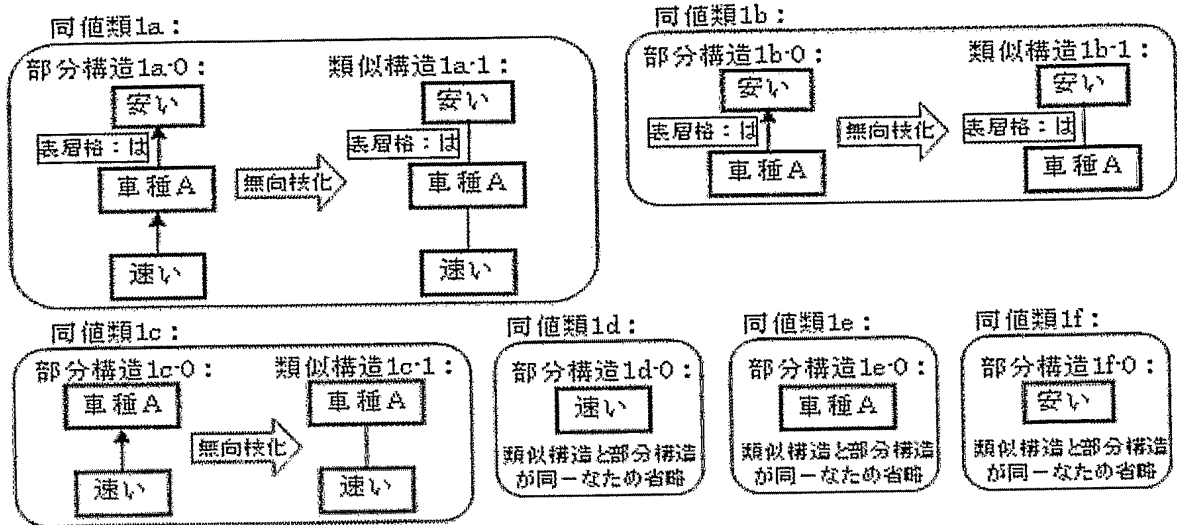
同値類2f:



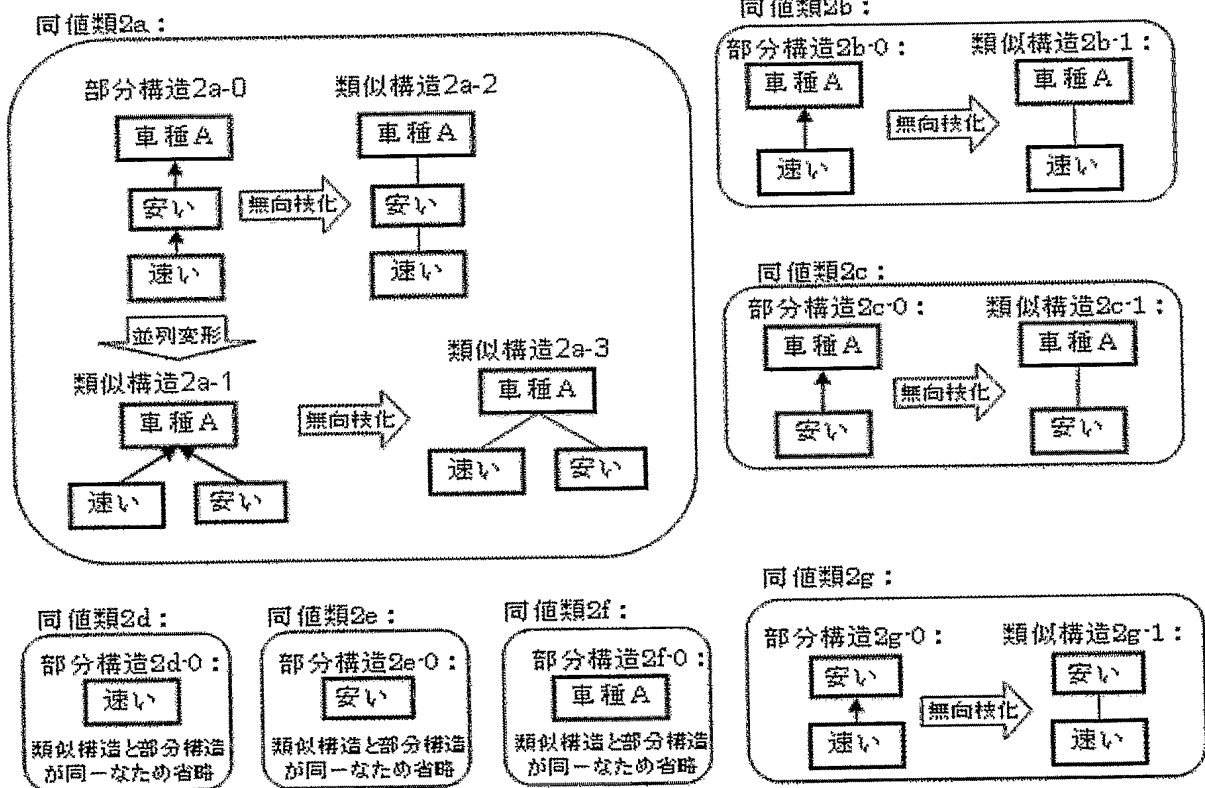
【図 2 2】



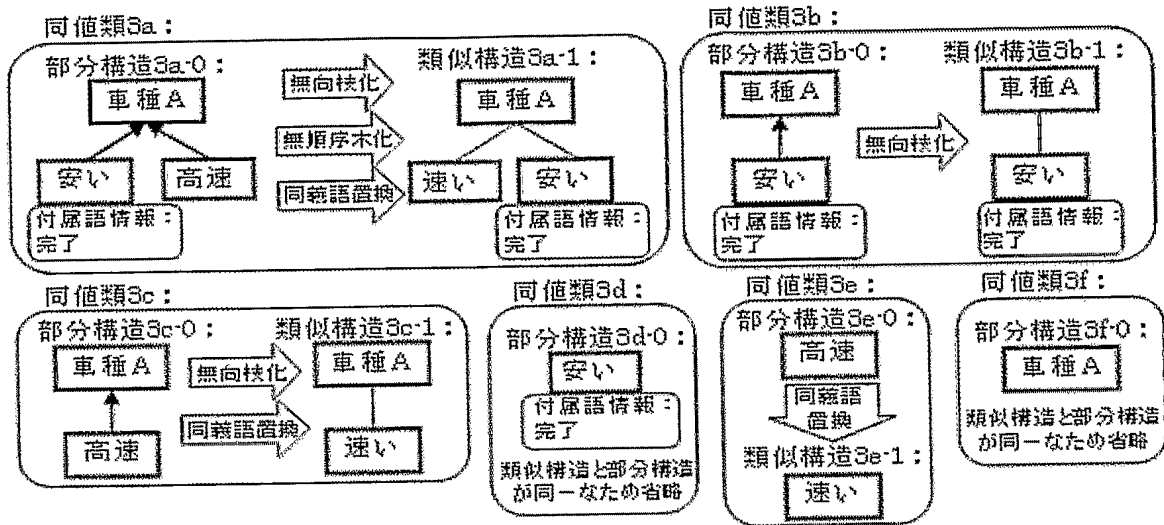
【図 2 3】



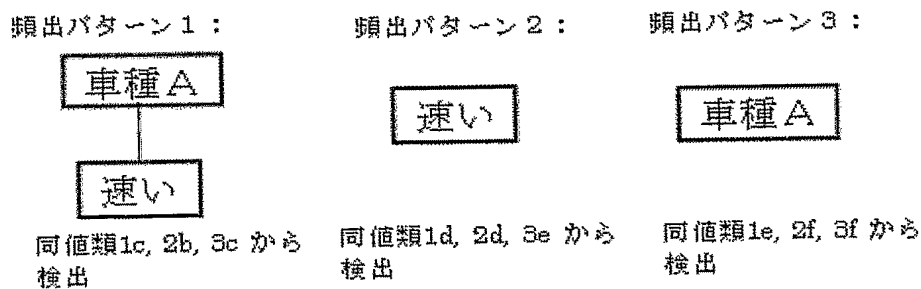
【図 2 4】



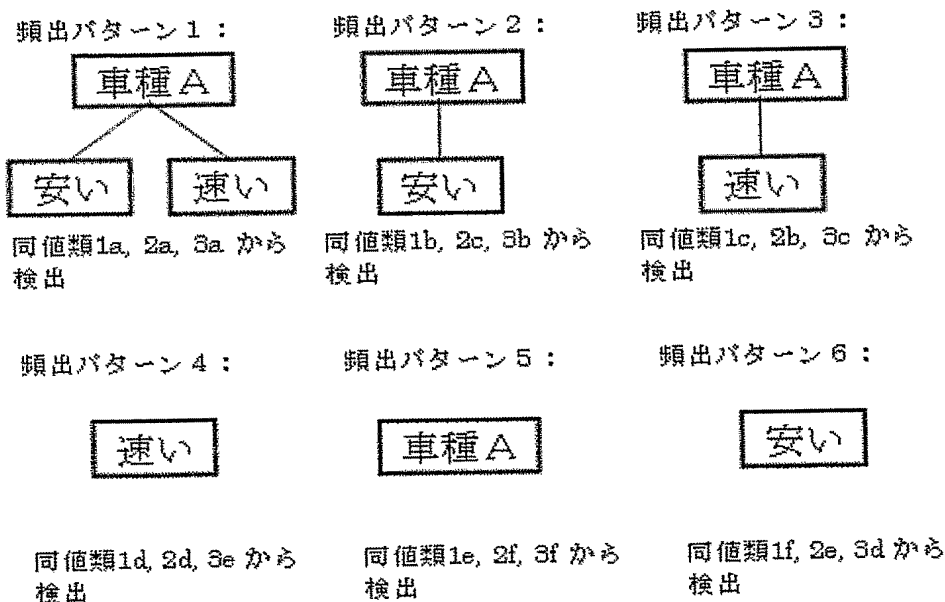
【図 25】



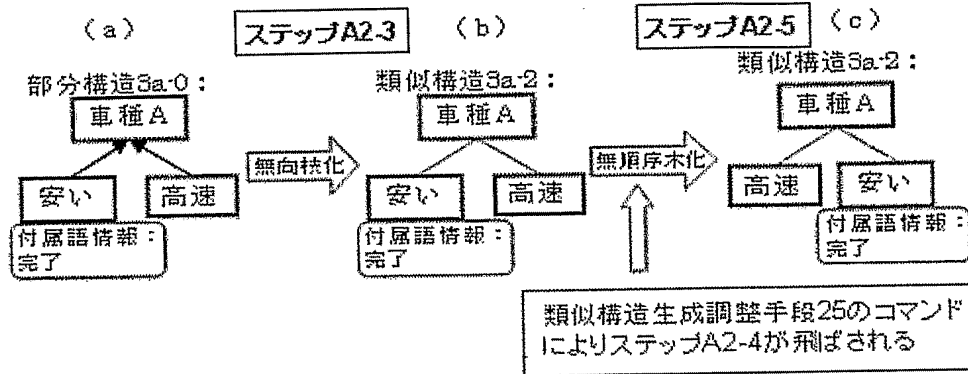
【図 26】



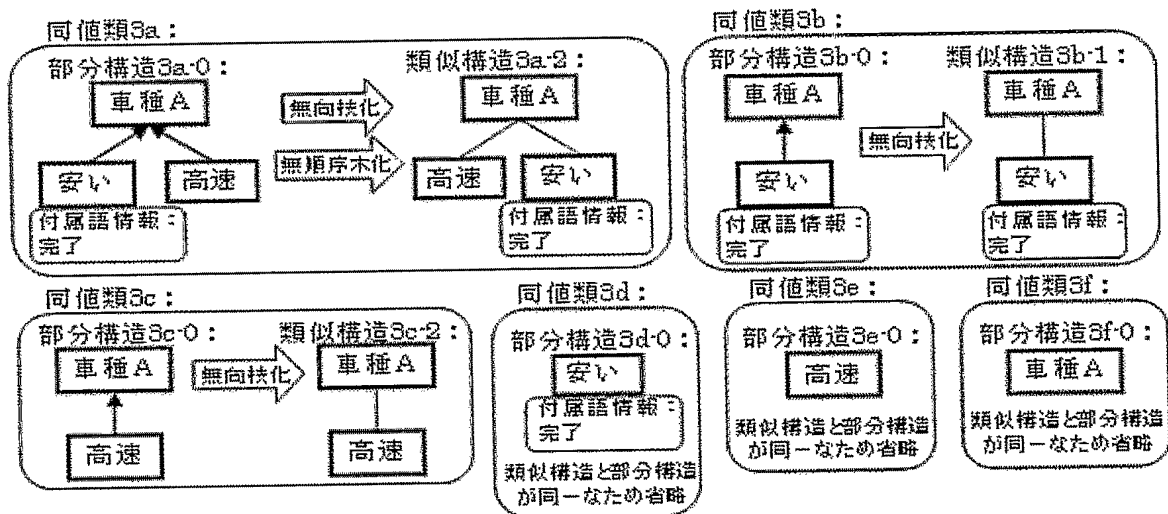
【図 27】



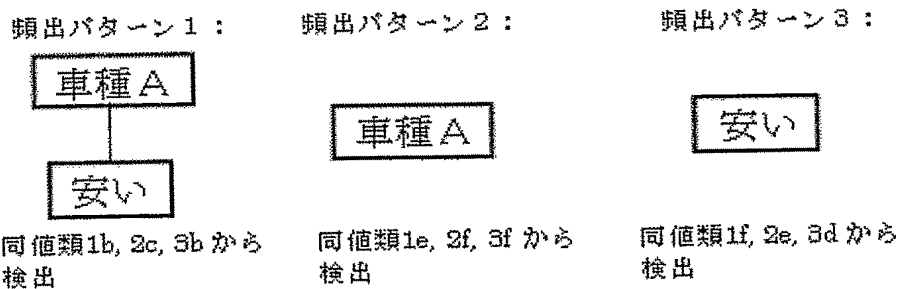
【図 28】



【図 29】



【図 30】



【書類名】 要約書**【要約】****【課題】**

意味的に類似した文構造を同一と判定しながら頻出パターンの検出を行うテキストマイニングを可能にする。さらに、どこまで類似した構造を同一と判定して頻出パターン検出を行うかを使用者が調整可能なテキストマイニングを可能にする装置、方法、プログラムの提供。

【解決手段】

言語解析手段 21 はテキスト DB 11 から読み込んだ各テキストの解析を行い解析結果として文構造を生成し、入力装置 6 は文構造の差異の種別毎に同一構造と判定するか否かを決定するための入力と属性値の種別毎に値の差異を無視するか否かを決定するための入力を受け、類似構造生成調整手段 25 は入力装置からの入力より文構造の差異の種別毎に同一構造と判定するか否かを指定する指定項目を生成し、類似構造判定調整手段 26 は入力装置 6 からの入力より属性値の種別毎に値の差異を無視するか否かを指定する指定項目を生成し、類似構造生成手段 22 は類似構造生成調整手段 25 からの指定項目に従い言語解析手段 21 が得た文構造の集合中の各構造の部分構造の類似構造を生成し生成した各類似構造を夫々の生成元の部分構造の同値類とし、頻出類似パターン検出手段 24 は類似構造判定調整手段 26 より与えられた指定項目に従い属性値を無視し類似構造生成手段 22 からの同値類の集合より頻出パターンを検出し出力装置 3 に出力する。

【選択図】

図 6



特願 2 0 0 4 - 0 7 9 0 7 7

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 4 2 3 7]

1. 変更年月日

1 9 9 0 年 8 月 2 9 日

[変更理由]

新規登録

住 所

東京都港区芝五丁目 7 番 1 号

氏 名

日本電気株式会社